OPEN ACCESS

KING'S College LONDON

King's Research Portal

*DOI:*
10.1109/ACCESS.2019.2928625

*Document Version*
Publisher's PDF, also known as Version of record

Link to publication record in King's Research Portal

*Citation for published version (APA):*
Zhao, Z., Bao, Z., Zhao, Y., Zhang, Z., Cummins, N., Ren, Z., & Schuller, B. (2019). Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition. *IEEE Access*, *7*(0), 97515-97525. https://doi.org/10.1109/ACCESS.2019.2928625

**Citing this paper**
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

**General rights**
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

**Take down policy**
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 04. May. 2023

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 19, 2019, accepted July 8, 2019, date of publication July 15, 2019, date of current version August 6, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2928625*

# Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition

**ZIPING ZHAO[1,3], ZHONGTIAN BAO[1], YIQIN ZHAO[1], ZIXING ZHANG[2] (Member, IEEE), NICHOLAS CUMMINS[3] (Member, IEEE), ZHAO REN[3], AND BJÖRN SCHULLER[2,3,4] (Fellow, IEEE)**

[1]College of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China
[2]GLAM – Group on Language, Audio & Music, Imperial College London, London SW7 2AZ, U.K.
[3]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany
[4]International Research Centre for Affective Intelligence, Tianjin Normal University, Tianjin 300387, China

Corresponding author: Björn Schuller (schuller@ieee.org)

This work was supported in part by the National Natural Science Foundation of China under Grant 61702370, in part by the Key Program of the Natural Science Foundation of Tianjin under Grant 18JCZDJC36300, in part by the Open Projects Program of the National Laboratory of Pattern Recognition, in part by the Senior Visiting Scholar Program of Tianjin Normal University, and in part by the European Union's Horizon 2020 Research And Innovation Programme under Grant 826506 (sustAGE).

**ABSTRACT** The automatic detection of an emotional state from human speech, which plays a crucial role in the area of human–machine interaction, has consistently been shown to be a difficult task for machine learning algorithms. Previous work on emotion recognition has mostly focused on the extraction of carefully hand-crafted and highly engineered features. Results from these works have demonstrated the importance of discriminative spatio-temporal features to model the continual evolutions of different emotions. Recently, spectrogram representations of emotional speech have achieved competitive performance for automatic speech emotion recognition (SER). How machine learning algorithms learn the effective compositional spatio-temporal dynamics for SER has been a fundamental problem of deep representations, herein denoted as deep spectrum representations. In this paper, we develop a model to alleviate this limitation by leveraging a parallel combination of attention-based bidirectional long short-term memory recurrent neural networks with attention-based fully convolutional networks (FCN). The extensive experiments were undertaken on the interactive emotional dyadic motion capture (IEMOCAP) and FAU aibo emotion corpus (FAU-AEC) to highlight the effectiveness of our approach. The experimental results indicate that deep spectrum representations extracted from the proposed model are well-suited to the task of SER, achieving a WA of 68.1 % and a UA of 67.0 % on IEMOCAP, and 45.4% for UA on FAU-AEC dataset. Key results indicate that the extracted deep representations combined with a linear support vector classifier are comparable in performance with eGeMAPS and COMPARE, two standard acoustic feature representations.

**INDEX TERMS** Speech emotion recognition, bidirectional long short-term memory, fully convolutional networks, attention mechanism, spectrogram representation.

## I. INTRODUCTION

Automatic emotion recognition from speech signals, aiming at the identification of our basic emotional states using machine learning, remains a difficult task. A major challenge currently being faced by researchers is how best to extract discriminative, robust, and affect-salient features that represent the acoustic contents of speech signals. Many previous research efforts have investigated several hand-crafted acoustic features for the task of speech emotion recognition (SER), such as prosodic features (e. g. , pitch, energy, zero-crossings), spectral features (e. g. , linear predictor coefficients (LPC), linear predictor cepstral coefficients (LPCC), mel-frequency cepstral coefficients (MFCC), and non-linear

The associate editor coordinating the review of this manuscript and approving it for publication was Haishuai Wang.

 This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/ 

features such as the Teager-energy-operator (TEO). More recently, with the increased use of neural networks for SER tasks, mel-scale filterbank spectrograms are now widely used as an input feature. Deep spectrum representations, which are features automatically extracted from speech spectrogram images using deep learning models, have produced promising results in the fields of SER [1] and other speech and audio related applications [1]–[3].

Inspired by their performance in visual recognition tasks [4], recent SER approaches such as deep spectrum have incorporated convolutional neural networks (CNNs) to extract features from spectrograms. CNNs are exceptionally good at capturing high-level representations in a spatial domain. Recently, fully convolutional networks (FCNs) [5] have been proposed as a variant of CNNs. A major advantage of FCNs is that they can handle inputs of variable sizes; based on this property, they have achieved state-of-the-art performance in time-series based classification tasks [6], [7].

However, a drawback of FCNs is that they are not primarily tailored for learning temporal features. In this regard, *recurrent neural networks with long short-term memory* (LSTM-RNNs) offer the advantage of being suitable to model temporal dependencies between sequences [8], and as a result are widely used in SER [9], [10]. The approach proposed herein aims to leverage the inherent strengths of the two aforementioned models. The framework combines, in a parallel manner, FCNs and LSTM-RNNS, specifically bidirectional LSTM-RNNs (BLSTM-RNNs), to learn effective compositional spatio-temporal dynamics from spectrograms for the SER task.

In addition to learning useful spatio-temporal features, it is also important to select the emotionally salient sections of an input signal to improve SER performance further [11]. The use of *attention mechanisms* in RNN and CNN-based models has frequently been demonstrated as a useful tool to encourage a model to more heavily weight specific regions of an input sequence or image [12]. Attention mechanisms have also been effectively applied in SER [11], [13]–[15].

Motivated by the above analysis, and following on from our previous preliminary work [10], [16], we propose the Attention-BLSTM-FCN model, a spatio-temporal spectrogram-based approach which leverages attention-based BLSTM-RNNs (Attention-BLSTM-RNNs) and attention-based FCNs in parallel for SER. An advantage of the Attention-BLSTM-FCN model is that it enables the model to capture both temporal and frequency dependence in the spectrogram of the speech, relying on FCNs to extract representations from the spectrogram and modelling the temporal dynamics using a BLSTM network. In order to focus on feature extraction in the emotionally salient parts of an utterance, we investigate the benefits of including attention-based architectures in the model. A concatenation operation is employed to take advantage of the complementary features extracted from BLSTM and FCN, and the learnt representations are then fed into a deep neural network (DNN) to predict the emotion of the input utterance.

The main contributions of this article are, therefore, as follows: i) we propose a novel framework to fuse both spatial and temporal representations for SER by leveraging attention-based FCNs with attention-based BLSTM-RNNs, an approach capable of automatically learning feature representations and modeling the temporal dependencies; ii) following the recent success of applying deep learning methods directly to spectrograms, enhanced deep spectrum representations are derived from forwarding spectrograms through the Attention-BLSTM-FCN model; and iii) the proposed method can be easily adapted to enhance existing state-of-the-art methods. To the best of the authors' knowledge, this is the first work in the literature that applies the Attention-BLSTM-FCN model to learn enhanced deep-spectrum representations for SER.

## II. RELATED WORK

SER is a highly active research field, with many novel approaches being proposed and investigated over the past decade. With the increase of available data and computational power, deep learning methods are rapidly becoming the predominant approach [17], [17]–[19]. In particular, many recent studies have explored leveraging deep neural networks as feature extractors to learn discriminative representation [20]. Due to their success in many visual recognition tasks, CNNs are being widely used in feature representation learning in various speech analysis tasks. For example, Huang et al. used spectrograms of speech together with a CNN to perform SER [21], and similar work is presented in [22], in which a CNN was employed to learn affect-salient features from spectrograms.

Nowadays, extracting spectrograms from audio clips and extracting deep spectrum representations by feeding them through a deep CNN has become a new research trend [1], [2], [23]–[25]. Furthermore, deep spectrum representations benefit from the advantage of transfer learning, as they are formed by passing spectrograms through pre-trained image classification deep CNNs such as AlexNet [26] or VGG [27]. Deep spectrum representations have been shown to produce suitable salient features which achieve state-of-the-art performance in a range of speech-related recognition tasks including SER [1].

Additionally, given that context information is crucial for detecting emotional states, RNN paradigms are widely used in SER to exploit the temporal information inherent in speech signals. LSTM-RNNs, in particular, are frequently employed in SER tasks [9], [11], [28]–[30].

Inspired by the success of CNNs and RNNs, there has been an increasing interest in incorporating both into a single architecture. For example, in [31], the Convolutional Long Short-Term Memory Deep Neural Networks (CLDNN) model was proposed for speech recognition. The developed model consisted of convolutional layers, LSTM gated recurrent layers, and fully connected (FC) layers. More recently, end-to-end network architectures have emerged as a promising network structure. These can automatically extract

representations directly from *raw* (unprocessed) data, rather than manually extracting hand-crafted features.

The SER approach proposed in [9] jointly exploited a CNN to automatically extract suitable representations from raw audio signals and an LSTM-RNN to capture the temporal information. A similar framework was proposed in [32] for the related task of speech-based depression detection. In [33], a specially designed neural network structure that accepts variable-length speech was proposed for SER. This approach combines CNN-based deep spectrogram representations with an RNN to handle the variable-length speech segments.

Similar to the Attention-BLSTM-FCN model developed in this paper, a parallel combination of LSTM and the CNN neural network framework has been explored for acoustic scene classification [34]. The results presented in [34] demonstrate that the LSTM model extracted key sequential information from consecutive audio features and the CNN model learnt salient spectro-temporal locality from spectrogram images.

Attention-based RNNs have begun to be widely used across a range of machine learning tasks. For example, they have been successfully applied in tasks such as speech recognition [15], natural language processing (NLP) [35], [36] and SER [10], [11], [30]. Similarly, attention mechanisms have also been exploited for CNNs for NLP tasks [37], audio-related classification tasks [24] and SER [6], [13], [14].

In summary, while there is a range of work in the literature focusing on feeding spectrograms into CNNs for speech-based recognition tasks, very little research has been undertaken to explore attention-based FCNs and attention-based LSTM-RNNs as mechanisms for extracting emotionally salient information from spectrograms.

## III. PROPOSED METHODOLOGY

In our proposed Attention-BLSTM-FCN model (cf. Fig. III), the Mel-spectrograms are fed into two parallel networks, namely an Attention-BLSTM and an Attention-FCN. We then concatenate the network outputs to form a new feature sequence. The Attention-BLSTM layers extract sequential information from the spectrograms, while the Attention-FCN layers extract spatial information. Fusion of the concurrently extracted and complementary features forms a joint spatio-temporal feature vector.

### A. SPECTROGRAM GENERATION

The first step in our proposed system is the extraction of the mel-spectrograms. Spectrograms are a time-frequency visual representation of a signal produced by a short-time Fourier transform (STFT) [38]. In the presented work, we used the *librosa*[1] framework to first resample the audio signals to 16 kHz, and then transform them to spectrograms utilizing the STFT implemented with a Hamming window function with a frame length of 25*ms* at a rate of 10*ms*. Following this, we mapped the STFT matrices into their magnitude squared via:

$$X_i(f, m) = |STFT\{x_i\}(f, m)|^2, \tag{1}$$

where $x_i$ is an utterance signal, $f$ stands for frequency and $m$ for window position. Finally, we generate the mel-spectrograms by scaling the $f$ hertz signal into $m$ mel-scaled bands via:

$$m = 2595 \log_{10}(1 + \frac{f}{700}). \tag{2}$$

Mel-frequency spacing approximates that of the human cochlea, and thus the resulting mel-spectrograms reflect the relative importance of different frequency bands as perceived by the human ea [39].

### B. ATTENTION-BASED BIDIRECTIONAL LONG SHORT-TERM MEMORY NETWORKS

Our proposed system includes the use of attention mechanisms, together with BLSTM in order to focus feature learning onto the salient regions of a sequence. The so-called Attention-Based BLSTM-RNN unit contains four components:

1) The input layer: the spectrogram is fed into the model.
2) An LSTM layer: utilizes a BLSTM to extract high-level representations from step (1).
3) An attention layer: produces a weight vector, and merges frame-level features from each time step into an utterance-level feature vector by multiplying it with the weight vector.
4) The output layer: outputs the resulting utterance-level feature representation.

We describe the LSTM and attention layers below in the following.

#### 1) BIDIRECTIONAL LONG SHORT-TERM MEMORY NETWORKS

As LSTM units solve the issue of vanishing and exploding gradients in RNN training [8], they are, usually, employed as the basic unit in RNN. An LSTM-RNN can, therefore, model long-range dynamic dependencies while avoiding issues relating to vanishing or exploding gradients during training. A standard LSTM can, however, only process sequential data in one direction [40], hence the BLSTM-RNN has been proposed to overcome this limitation. In a BLSTM-RNN, the input is processed both in the standard order and reversed order, allowing the network to combine future and past information at every time step.

A BLSTM component comprises two LSTM layers processing the input separately to produce $\overrightarrow{h}$, $\overrightarrow{c}$, the hidden states and the cell states of an LSTM processing the input in the forward direction, and $\overleftarrow{h}$, $\overleftarrow{c}$, the hidden states and cell states of an LSTM processing the input in reversed order. Both $\overrightarrow{h}$, and $\overleftarrow{h}$, are then combined using:

$$y_t = W_{\overrightarrow{h}y}\overrightarrow{h}\,t + W_{\overleftarrow{h}y}\overleftarrow{h}\,t + b_y, \tag{3}$$

to produce the output sequence of the BLSTM layer.

[1] https://github.com/librosa/librosa

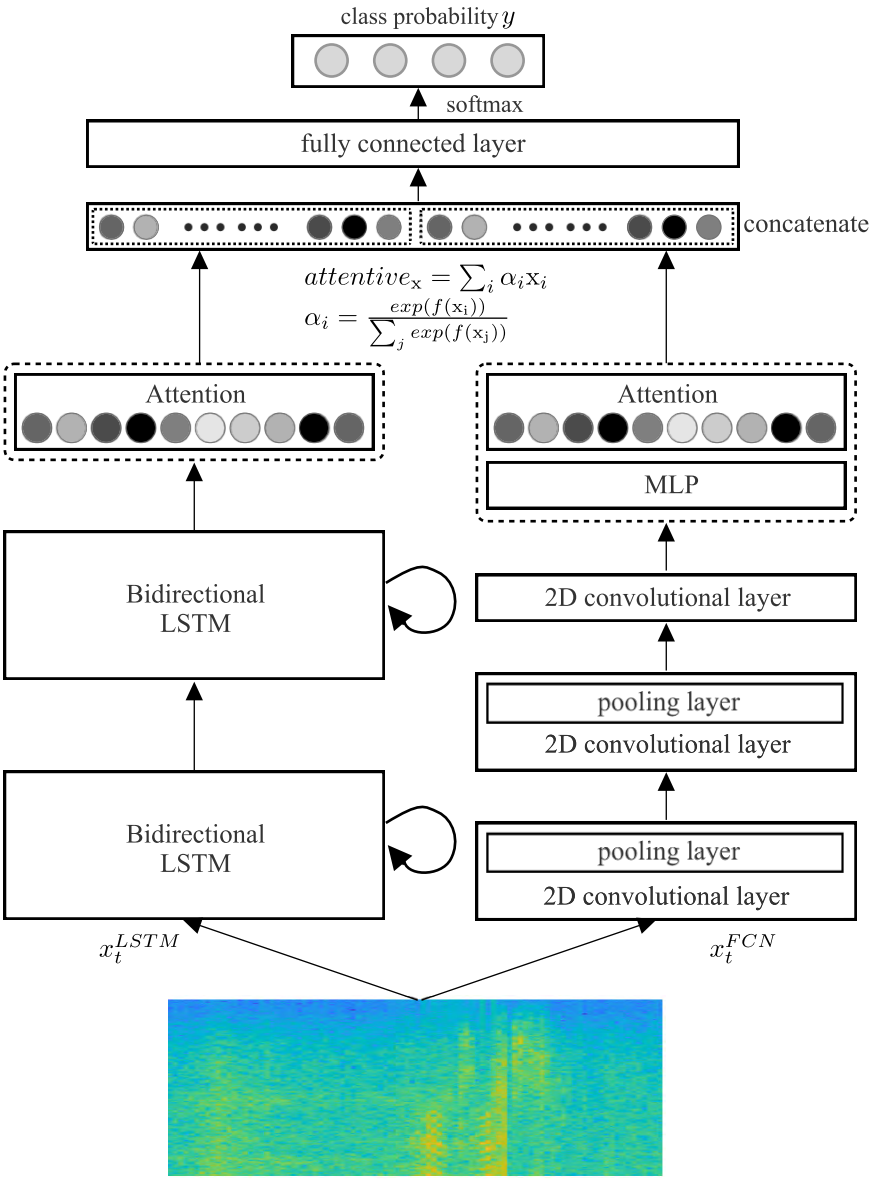



**FIGURE 1.** An overview of our proposed Attention-BLSTM-FCN model, in which spectrograms are fed into two parallel networks, an Attention-BLSTM to extract temporal information, and an Attention-FCN to extract spatial information. The fusion of the outputs of these networks forms a joint spatial-temporal feature vector.

Note that it is also possible to use the cell states, instead of the hidden states, of the two LSTM layers in a BLSTM layer to produce the output sequence of the BLSTM layer:

$$y_t = W_{\overrightarrow{c}y}\overrightarrow{c}\,t + W_{\overleftarrow{c}y}\overleftarrow{c}\,t + b_y. \tag{4}$$

#### 2) ATTENTION LAYER

In this layer, a 1D attention module is built on top of the BLSTM layer. To determine the attention weights $\alpha_i$, we calculate each vector entry $x_i$ in a sequence of inputs $x$, as follows:

$$\alpha_i = \frac{exp(f(x_i))}{\sum_j exp(f(x_j))}, \tag{5}$$

in which $f(x)$ denotes the scoring function. We use $f(x) = W^T x$ for $f(x)$, in which $W$ is the trainable parameter, as a linear scoring function.

The output of the attention layer is then the weighted sum of the input sequence, defined as $attentive_x$:

$$attentive_x = \sum_i \alpha_i x_i. \tag{6}$$

### C. ATTENTION POOLING BASED FULLY CONVOLUTIONAL NETWORKS

Our proposed system also includes the use of attention mechanisms, together with FCN in order to focus feature learning

onto more emotion-relevant time-frequency regions of the mel-spectrograms of speech.

#### 1) FULLY CONVOLUTIONAL NETWORKS

Similar to a conventional CNN, the FCN structure only consists of convolutional layers, and hence the local feature structures are effectively preserved with a relatively small number of weights. Meanwhile, the FCN structure also provides advantages by allowing the networks to model the temporal and harmonic structure of audio signals [41]. Given these benefits, we use spatial convolutional neural networks with an FCN-like structure for our deep spectrum features extraction.

In this work, the output of our FCN is a three-dimensional array of size $F \times T \times C$, where the $F$ and $T$ stands for the frequency and time domains of the spectrogram and $C$ for channel size. We consider the output as a variable-length grid of $L$ elements, $L = F \times T$. In set $A$, each of the elements is a C-dimensional vector corresponding to a region of speech spectrogram, represented as $\alpha_i$.

$$A = \{\alpha_1, \ldots, \alpha_L\}, \quad \alpha_i \in \mathbb{R}^C. \tag{7}$$

In this work, we employ a 3-layer FCN which contains three convolutional layers and three max-pooling layers. The network takes a log-amplitude mel-spectrogram sized $40 \times 500$ as input and predicts a 128-dimensional output vector. As the FCN is performing feature extraction, its final output comes from the attention pooling [42], which reduces the number of parameters of the network.

#### 2) ATTENTION POOLING METHOD

As not all time-frequency units will contribute equally to the emotional state associated with an utterance, we, therefore, adopt an attention mechanism similar to [6]. We place it on top of the FCN to help the network pay more attention to specific time-frequency regions of the input spectrogram. We realize the attention module as follows. First, the annotation $a_i$ is fed as input to obtain a new representation of $a_i$ through a multilayer-perceptron (MLP) layer employing tanh as the non-linear activation function:

$$e_i = u^T tanh(Wa_i + b). \tag{8}$$

Next, we calculate the importance weight, $e_i$, of the $a_i$ by the inner product between this new vector and the learnable vector $u$. After this, the normalized importance weight $\alpha_i$ is calculated using the softmax function:

$$\alpha_i = \frac{exp(\lambda e_i)}{\sum_{k=1}^{L} exp(\lambda e_k)}. \tag{9}$$

In this equation, $\lambda$ is a scale factor which controls the uniformity of the importance weights of the annotation vectors. $\lambda$ ranges between 0 and 1. If $\lambda = 1$, the scaled-softmax becomes the commonly used softmax function. If $\lambda = 0$, the importance weights will be a uniform distribution on the set $A$, which means all the time-frequency units have the same importance weights for the final utterance emotion vector. In this work, we set $\lambda = 0.3$ according to the performance on the validation set [6]. Finally, the utterance emotion vector $c$ is computed as the weighted sum of set $A$ with importance weights:

$$c = \sum_{i=1}^{L} \alpha_i a_i. \tag{10}$$

**TABLE 1. Instance distribution over four emotion classes for the IEMOCAP Dataset.**

| Session | Neutral | Happy | Sad | Angry | Total |
|---|---|---|---|---|---|
| 1 | 223 | 132 | 104 | 62 | 521 |
| 2 | 217 | 191 | 100 | 22 | 530 |
| 3 | 198 | 149 | 190 | 90 | 627 |
| 4 | 174 | 195 | 81 | 84 | 534 |
| 5 | 287 | 280 | 133 | 31 | 731 |
| Sum | 1 099 | 947 | 608 | 289 | 2 943 |

## IV. EXPERIMENTS AND RESULTS

In this section, we provide key details relating to the experimental setup, our experiments, and the results of our analysis.

### A. DATASET DESCRIPTION

IEMOCAP consists of audio-visual data with transcriptions from recordings of dialogues between two professional actors, over 5 sessions, with the corpus divided into two parts: *improvise* and *script* [43]. In our experiments, we only focus on the improvised sessions. Adopting the methodology of previous works, we used a leave-one-session-out strategy. In each training process, 8 speakers from 4 sessions were used as training data, and the remaining session was separated into two parts: one being regarded as validation data and the other as test data. It is also worth noting here that the data distribution of each emotion class is heavily imbalanced. As in [44], we, therefore, merge the happy and excited utterances into the happy class since they are close in emotion. Four emotion categories are, therefore, employed in the training and evaluation: *angry*, *happy*, *sad*, and *neutral* (cf. Table 1).

FAU Aibo Emotion Corpus (FAU-AEC), on the other hand, is composed of spontaneous and emotional German speech samples [45]. The corpus contains 9.2 hours of German speech from a total of 51 children interacting with Sony's pet robot Aibo at two different schools. As per [46], we used 9 959 utterances from 26 children (13 males and 13 females from the Ohm School) as the training set and 8 257 utterances from 25 children (8 males and 17 females from the Montessori School) as the test set.[2] In this study, we concentrated on the five-class problem with the emotion categories of *anger*, *emphatic*, *neutral*, *positive*, and *rest* (cf. Table 2).

[2]We will provide a URL for a document with details on partitions and seeds upon acceptance.

**TABLE 2. Instance distribution over five emotion classes for the FAU Aibo Emotion Corpus.**

| | Angry | Emphatic | Neutral | Positive | Rest | Total |
|---|---|---|---|---|---|---|
| Train | 881 | 2 093 | 5 590 | 674 | 721 | 9 959 |
| Test | 661 | 1 508 | 5 377 | 215 | 546 | 8 257 |
| Total | 1 492 | 3 601 | 10 967 | 889 | 1 267 | 18 216 |

### B. EXPERIMENT SETUP AND EVALUATION METRICS

The proposed Attention-BLSTM-FCN model has many hyperparameters, a proportion of these being tuned based on the recommendations from previous works which utilized the same database [10], [16]. In order to identify the optimal model, we optimized 15 hyperparameters: window size, convolutional kernel size, pooling size, stride on convolutional layer, initial number of filters and neurons, learning rate, the number of convolutional/pooling/fully connected layers, type of activation function, optimization algorithm, dropout on convolutional and fully connected layers, and frequency resolutions of the input spectrogram. The details on these hyper-parameters are given below:

1) We set the window size to 25 ms (window sizes between 15 ms to 200 ms were tested) and window shift is set to 10 ms
2) The BLSTM contained 128 $\times$ 2 nodes. We also tested BLSTMs of 64 $\times$ 2 nodes however we observed an accuracy drop of 1-3 %
3) Our Mel-spectrograms were formed using 40 Mel bands (30, 60, 80, and 100 bands were also tested)
4) The optimal FCN topology was found to be 3 layers (we tested 2-5 layers), similarly, the best topology for BLSTM is found to be 2 layers (we tested 1-3 layers)
5) The FCN filters are set to 64, 128, 128 (each layer was tested from 8 to 256). Stride for the CNN layers was set as (1, 1).
6) A dropout layer, batch normalization techniques, and ReLU activation functions are applied to prevent overfitting.
7) The Adam optimizer with a learning rate of $10^{-3}$, and a decay of $10^{-6}$ is used for training.
8) All models were implemented by the *TensorFlow*[3] framework.
9) All models were trained with a maximum epoch of 100 and batch size of 100 with dropout regularization utilized to prevent overfitting.

To evaluate the performance of the proposed framework, we conducted several experiments. First, in order to investigate the influence of spatial and temporal information, we built our FCNs, attention-FCN, and attention-BLSTM models as described above. A comparison of FCNs, attention-FCN, attention-LSTM, attention-BLSTM, as well as our proposed model was performed. We then evaluated the performance of the standard spectrogram with different spectrogram resolutions based on the Attention-BSLTM-FCN model. Note, resolution is an important decision when generating models that rely on spectrograms. The work presented in [47] reveals the performance differences among different frequency resolutions of the input spectrogram. To this end, the Attention-BLSTM-FCN model was re-trained using either a 30-band, a 40-band, 60-band, 80-band, or 100-band Mel-spectrogram. Moreover, spectrograms represent a 2D representation of audio signals. On the one hand, changes in the Mel-scale represent a scale effect on the vertical direction. On the other hand, the horizontal scale of each data point is influenced by (temporal) window length (cf. Figure 2). We therefore also tested the effect of varying window sizes between 15 ms to 200 ms.

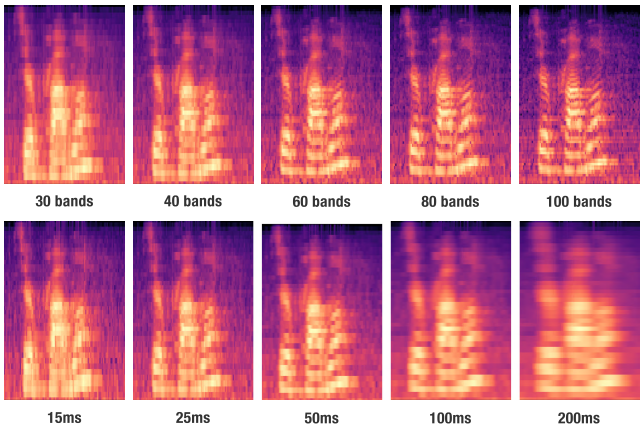



**FIGURE 2. A Visual comparison of different Mel scaled frequency resolutions and different STFT window lengths of 2-seconds-long spectrogram fragments.**

Thirdly, we compared the effectiveness of the deep spectrum representations extracted from the Attention-BSLTM-FCN model with two commonly used SER feature representations: *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [48] and *Interspeech Computational Paralinguistics Challenge (*ComParE*) features set*. In order to do so, we first extracted the eGeMAPS and ComParE low level features with the openSMILE toolkit [49]. Due to the high dimension of the ComParE feature set, we performed the PCA technique on the training set to reduce the feature size by selecting top 150 components which explained >95 % variances of the original features. We then applied functionals (max, min, range, mean, and standard-deviation) on the two feature representations independently for each combination of speaker and feature independently. Finally, all the feature representations are fed into a linear *support vector machine* (SVM) implemented using the *scikit-learn*[4] toolbox, and trained via stochastic gradient descent.

Finally, in order to show the effectiveness of our approach, we compared the performance with systems based on pre-trained CNNs, namely 'AlexNet' [26], 'VGG16', and 'VGG19' [27]. We obtained the pre-trained 'AlexNet' network from MATLAB R2017a3, and 'VGG16' and

[3]https://www.tensorflow.org

[4]http://scikit-learn.org

**TABLE 3. Configurations of 'AlexNet', 'VGG16', and 'VGG19'.**

| AlexNet | VGG16 | VGG19 |
|---|---|---|
| input: RGB image | | |
| 1×conv11-96 | 2×conv3-64 | 2×conv3-64 |
| maxpooling | | |
| 1×conv5-256 | 2×conv3-128 | 2×conv3-128 |
| maxpooling | | |
| 1×conv3-384 | 3×conv3-256 | 4×conv3-256 |
| | maxpooling | |
| 1×conv3-384 | 3×conv3-512 | 4×conv3-512 |
| | maxpooling | |
| 1×conv3-256 | 3×conv3-512 | 4×conv3-512 |
| maxpooling | | |
| fully connected layer *fc*6-4096 | | |
| fully connected layer *fc*7-4096 | | |
| fully connected layer *fc*-1000 | | |
| output: soft-max | | |

**TABLE 4. Class weights for data balance when using the FAU Aibo Emotion Corpus.**

| | Angry | Emphatic | Neutral | Positive | Rest |
|---|---|---|---|---|---|
| weight | 1.1 | 0.5 | 0.2 | 1.5 | 1.4 |

'VGG19' from MatConvNet [50]. Then, we exploited the Mel-spectrograms as the input for these three pre-trained CNNs and extracted the deep representations from the activations on the second fully connected layer (fc7) as feature vectors (cf. Table 3). The feature representations extracted by the three pre-trained CNNs and Attention-BLSTM-FCN were fed into the linear SVM.

As evaluation measures, we employ the standard evaluation criteria used on the IEMOCAP and FAU-AEC dataset. For IEMOCAP, we used both unweighted and weighted accuracies (UA and WA respectively) as the evaluation metric, while for FAU-AEC, we use only unweighted accuracy (UA) as the evaluating measure as this database is extremely unbalanced. Furthermore, in order to tackle the problem of unbalanced data, we apply class weights during training (cf. Table 4) identified using:

$$r_k = \frac{N}{N_k} \propto \frac{1}{N_k}, \tag{11}$$

where N is the total number of training examples, and $N_k$ is the number of the training examples of each class [51].

### C. RESULTS

A comparison shows that the Attention-BLSTM-FCN model achieves the best performance. It can be seen that the proposed approach outperforms previous works on the IEMOCAP and FAU-AEC datasets (cf. Table 5). Our highest UA and WA achieved on IEMOCAP were 68.1 %, and 67.0 %, respectively. This represents a significant improvement over the baseline FCN model ($p < .05$ in a one-tailed z-test). The same system set-up also achieved the best UAR, 45.4 %, on FAU-AEC. Again, this represents a significant improvement over the baseline FCN ($p < .05$ in a one-tailed z-test).

**TABLE 5. Performance comparison between the proposed Attention-BLSTM-FCN with other models on the IEMOCAP and FAU Aibo Emotion corpus.**

| Models | IEMOCAP | | FAU-AEC |
|---|---|---|---|
| | WA[%] | UA[%] | UA[%] |
| DNN+ELM [52], [53] | 57.9 | 52.1 | – |
| RNN+ELM [53] | 62.9 | 63.9 | – |
| CNN+LSTM [54] | 67.3 | 62.0 | – |
| Attention+FCN [6] | 67.9 | 57.3 | – |
| Anchor models [55] | – | – | 44.0 |
| DBN+HMM [56] | – | – | 45.0 |
| Skew-Robust Neual Networks [57] | – | – | 45.3 |
| FCNs | 64.5 | 62.0 | 42.1 |
| Attention-FCN | 63.0 | 64.1 | 42.9 |
| Attention-LSTM | 62.8 | 60.9 | 42.6 |
| Attention-BLSTM | 65.5 | 64.8 | 43.5 |
| Attention- One-layer BLSTM-FCN | 66.3 | 65.8 | 44.2 |
| Attention- Two-layer BLSTM-FCN | **68.1** | **67.0** | **45.4** |
| Attention- Three-layer BLSTM-FCN | 66.7 | 66.2 | 45.0 |

Note: for IEMOCAP we use both unweighted and weighted accuracies (UA and WA respectively) as the evaluation metric, while for FAU-AEC, we only adopt the UA as the evaluating measure since the FAU-AEC is extremely unbalanced.

**TABLE 6. Performance comparison between different mel-bands on the IEMOCAP and FAU Aibo Emotion corpus.**

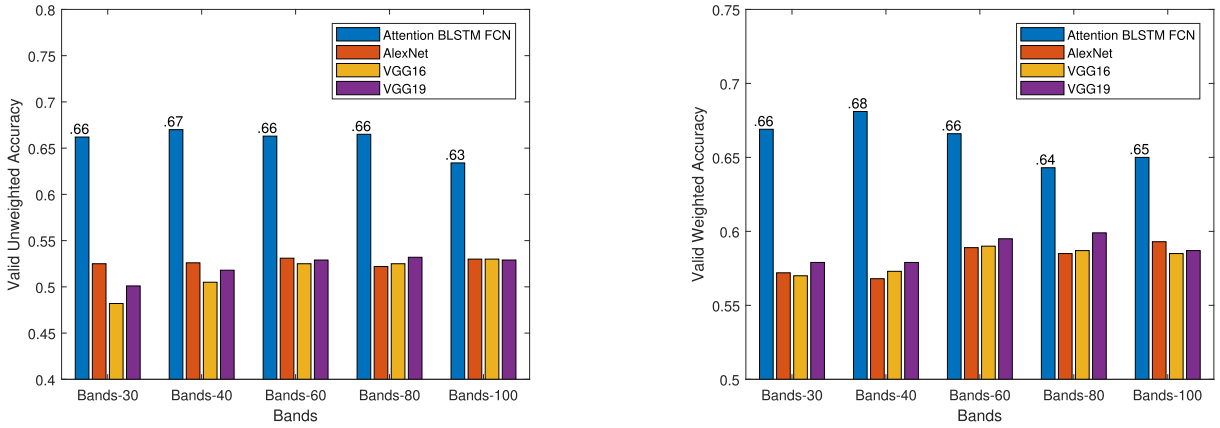
| Mel-bands | IEMOCAP | | FAU-AEC |
|---|---|---|---|
| | WA[%] | UA[%] | UA[%] |
| 30 | 66.9 | 66.2 | 42.9 |
| 40 | **68.1** | **67.0** | **45.4** |
| 60 | 66.6 | 66.3 | 43.0 |
| 80 | 64.3 | 66.5 | 42.7 |
| 100 | 65.0 | 63.4 | 41.0 |

**TABLE 7. Performance comparison of the deep spectrum features with eGeMAPS and ComParE feature sets.**

| Feature Set | IEMOCAP | | FAU-AEC |
|---|---|---|---|
| | WA[%] | UA[%] | UA[%] |
| eGeMAPS | 59.9 | 59.1 | 41.9 |
| ComParE | 63.1 | 61.9 | 37.7 |
| Deep Spectrum | **66.7** | **66.5** | **43.9** |

Our second experiment explores the difference in the frequency resolution in our system set-up (cf. Table 6) We observed that frequency plays an important role when extracting deep features. In this group of experiments, the best performances were 68.1 % (WA) and 67.0 % (UA) on IEMOCAP; 45.4% (UA) on FAU-AEC was achieved by the resolution of 40 Mel-bands.

When comparing our features with two standard acoustic features (cf. Table 7), we observed that the best UA (66.5 %) and WA (66.7 %) on IEMOCAP and the best UA (43.9 %) on FAU-AEC were achieved by the deep spectrum features extracted from our proposed model. This set-up yielded a significant improvement over the eGeMAPS ($p < 0.01$ in a one-tailed z-test) and C OMP ARE ($p < 0.01$ in a one-tailed z-test) feature sets. These comparisons indicate the promise of the deep spectrum features; further investigations are warranted to establish their suitability over a range of speech-related tasks.




(a) Unweighted Accuracy (b) Weighted Accuracy

**FIGURE 3.** Performance comparison between the proposed Attention-BLSTM-FCN with pre-trained models on the IEMOCAP dataset.

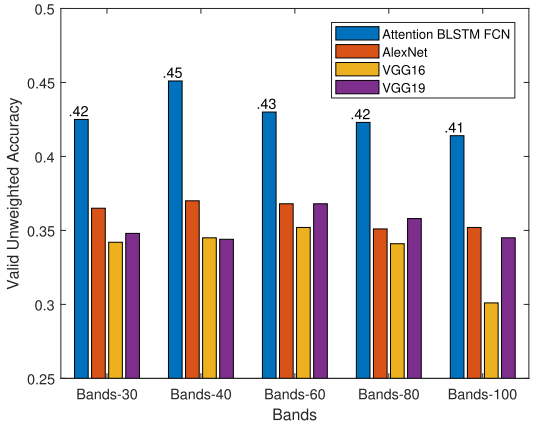



**FIGURE 4.** Performance comparison between the proposed Attention-BLSTM-FCN with pre-trained models on the FAU Aibo Emotion corpus.

Finally, when comparing the Attention-BLSTM-FCN model with more conventional deep spectrum approaches, the advantages of this framework can be clearly seen (cf. Fig. 3 and Fig. 4). Across the two data sets, and across the different mel-frequency resolutions, the Attention-BLSTM-FCN approach also yielded a significant improvement over deep spectrum representations extracted by AlexNet, 'VGG16' and 'VGG19' ($p < .01$ in a one-tailed z-test). Given the previous results showing the suitability of AlexNet, in particular, for deep spectrum feature extraction [1], [2], [10], [16], these results highlight the effectiveness of our proposed model for SER.

### D. DISCUSSION

From an overall experimental view point, the presented results demonstrate that our proposed model achieves notable performance improvements over the other, existing methods on IEMOCAP as well as the FAU-AEC. Furthermore, the proposed model outperforms both the baseline models and the individual application of attention-FCN and attention-BLSTM. These comparisons imply that it is crucial to use both spatial and temporal spectral information to boost speech emotion recognition and analysis. In terms of improved performance, it is clear that both the attention-FCN and the attention- BLSTM models complement each other. The consistently stronger performances of the Attention-BSLTM-FCN deep features compared to the other three deep pre-trained convolutional neural networks (cf. Fig. 3 and Fig. 4) support this hypothesis.

Our results also demonstrate that, on average, attention mechanisms can improve the prediction accuracy of the FCNs and BLSTM modules. We observed that the attention-FCN module did not result in a consistent improvement in WA over use of the FCN model alone when using the IEMOCAP dataset. In this regard, it is important to note that WA is highly dependent on the distribution of classes in the dataset. Therefore, we lend more importance to the UA; it better reflects the imbalanced distribution of the emotional classes. A comparison of the results for the proposed architecture with those for the eGeMAPS and ComParE feature sets indicate that it performs well as a feature extractor. It is worth noting that we did not perform any preprocessing on data and only used an SVM for classification. Additionally, the recognition results based on the deep spectrum representations derived from the proposed model outperformed the other two commonly used feature sets. These results add to the growing evidence in the literature that forwarding spectrogram representations through deep learning models produces salient features suitable for speech-related classification tasks.

We also observed that the frequency resolution of the input spectrogram is an important factor in determining the overall performance of the model (cf. Table 6). This effect is most likely due to the network learning some form of frequency discriminating function. Consistent with some ~~other~~ results in the literature [58], setting the frequency resolution to 40 mel bands yields better results than those with any

other value. This result contradicts those presented in [47], in which it was observed that using a higher number of mel frequency bands uniformly improved system accuracy. However, a reasonable explanation for this could be the difference in the models employed. Thus, the number of mel-bands required should potentially be treated as a hyperparameter and evaluated on a case-by-case basis.

Even though our results are more than encouraging, our approach has several limitations and a number of research directions should be considered for future research. A potential limitation of our proposed model is increased computations due to the generation of more trainable weights and hyperparameters. Moreover, further research needs to be conducted to confirm the robustness of our proposed model. Furthermore, we expect that the application of our approach to large datasets would show bigger improvements with respect to deep spectrum representations.

## V. CONCLUSION

We have proposed and developed a joint deep neural network architecture comprising a parallel combination of attention enhanced FCN and BLSTM networks to perform efficient SER from spectrograms. We trained an Attention-BLSTM-FCN model based on the spectrograms generated from the IEMOCAP and FAU-AEC datasets. The results of our experiments are highly promising, providing a new direction for consideration when performing emotion recognition.

In future work, we plan to further realize the potential of our proposed model and deep spectrum representations by establishing their suitability in other speech and acoustic recognition tasks.

## ACKNOWLEDGMENT

We thank Dr. Judith Dineley for her proof-reading work.

## REFERENCES

[1] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proc. 25nd ACM Int. Conf. Multimedia (ACMMM)*, Mountain View, CA, USA, Oct. 2017, pp. 478–484.

[2] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 3512–3516.

[3] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "Acoustic scene classification using convolutional neural networks," in *Proc. IEEE AASP Challenge Detection Classification Acoustic Scenes Events (DCASE)*, Budapest, Hungary, Jun. 2016, pp. 95–99.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and A. C. Berg, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for Semantic segmentation," in *Proc. 28th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Aug. 2015, pp. 3431–3440.

[6] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Honolulu, HI, USA, Nov. 2018, pp. 1771–1775.

[7] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[10] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition," in *Proc. INTERSPEECH*, Hyderabad, India, May 2018, pp. 272–276.

[11] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using Recurrent Neural Networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 2227–2231.

[12] C. Gorrostieta, R. Brutti, K. Taylor, A. Shapiro, J. Moran, A. Azarbayejani, and J. Kane, "Attention-based sequence classification for affect detection," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 506–510.

[13] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 1263–1267.

[14] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 3087–3091.

[15] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, Canada, Aug. 2015, pp. 577–585.

[16] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li, "Deep spectrum feature representations for speech emotion recognition," in *Proc. Joint Workshop 4th Workshop Affect. Social Multimedia Comput. 1st Multi-Modal Affect. Comput. Large-Scale Multimedia Data*, Seoul, Korea, Oct. 2018, pp. 27–33.

[17] H. Wang, Q. Zhang, J. Wu, S. Pan, and Y. Chen, "Time series feature learning with labeled and unlabeled data," *Pattern Recognit.*, vol. 89, pp. 55–66, May 2019.

[18] H. Wang, Z. Cui, Y. Chen, M. Avidan, A. B. Abdallah, and A. Kronzer, "Predicting hospital readmission via cost-sensitive deep learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 6, pp. 1968–1978, Dec. 2018.

[19] H. Wang, J. Wu, S. Pan, P. Zhang, and L. Chen, "Towards large-scale social networks with online diffusion provenance detection," *Comput. Netw.*, vol. 114, pp. 154–166, Feb. 2017.

[20] H. Wang, J. Wu, P. Zhang, and Y. Chen, "Learning shapelet patterns from network-based time series data," *IEEE Trans Ind. Informat.*, vol. 15, no. 7, pp. 1–14, Dec. 2018.

[21] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proc. 22nd ACM Int. Conf. Multimedia (ACMMM)*, Orlando, FL, USA, Nov. 2014, pp. 801–804.

[22] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[23] Z. Ren, V. Pandit, K. Qian, Z. Yang, Z. Zhang, and B. Schuller, "Deep sequential image features on acoustic scene classification," in *Proc. IEEE AASP Challenge Detection Classification Acoustic Scenes Events (DCASE)*, Munich, Germany, Dec. 2017, pp. 113–117.

[24] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley, and B. Schuller, "Attention-based Convolutional Neural Networks for acoustic scene classification," in *Proc. IEEE AASP Challenge Detection Classification Acoustic Scenes Events (DCASE)*, Surrey, U.K., Aug. 2018, 5 pages.

[25] Z. Ren, N. Cummins, V. Pandit, J. Han, K. Qian, and B. Schuller, "Learning image-based representations for heart sound classification," in *Proc. Int. Digital Health Conference (DH)*, Lyon, France, Apr. 2018, pp. 143–147.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 26th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1097–1105.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, Jul. 2015, p. 14.

[28] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008, pp. 597–600.

[29] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, San Antonio, TX, USA, Oct. 2017, pp. 190–195.
[30] C. W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition," in *Proc. INTERSPEECH*, San Francisco, CA, USA, 2016, pp. 1387–1391.
[31] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 4580–4584.
[32] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "DepAudioNet: An efficient deep model for audio based depression classification," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge (AVEC)*, Amsterdam, The Netherlands, Oct. 2016, pp. 35–42.
[33] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 3683–3687.
[34] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proc. IEEE AASP Challenge Detection Classification Acoustic Scenes Events (DCASE)*, Budapest, Hungary, Sep. 2016, pp. 11–15.
[35] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, Canada, 2015, pp. 2773–2781.
[36] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473
[37] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "ABCNN: Attention-based convolutional neural network for modeling sentence pairs," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 259–272, Jun. 2016.
[38] E. Sejdić, I. Djurović, and J. Jiang, "Time–frequency feature representation using energy concentration: An overview of recent advances," *Digit. Signal Process.*, vol. 19, no. 1, pp. 153–183, Jan. 2009.
[39] D. O'shaughnessy, *Speech Communication*. Boca Raton, FL, USA: Universities Press, 1987.
[40] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[41] K. Choi, G. Fazekas, and M. B. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proc. 17th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, New York, NY, USA, Jul. 2016, pp. 1–8.
[42] M. Lin, Q. Chen, and S. Yan, "Network in network," Dec. 2013, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400
[43] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, p. 335, Dec. 2008.
[44] R. Xia and Y. Liu, "DBN-ivector framework for acoustic emotion recognition," in *Proc. INTERSPEECH*, San Francisco, CA, USA, Sep. 2016, pp. 480–484.
[45] S. Steidl, *Automatic Classification of Emotions in Spontaneous Speech*. Berlin, Germany: Univ. Erlangen-Nuremberg Erlangen, 2009.
[46] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, U.K., 2009, pp. 312–315.
[47] K. J. Piczak, "The details that matter: Frequency resolution of spectrograms in acoustic scene classification," in *Proc. IEEE AASP Challenge Detection Classification Acoustic Scenes Events (DCASE)*, Munich, Germany, Jul. 2017, pp. 103–107.
[48] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andrè, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Jun. 2016.
[49] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia (ACMMM)*, Barcelona, Spain, Oct. 2013, pp. 835–838.
[50] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia (ACMMM)*, Oct. 2015, pp. 689–692.
[51] P.-W. Hsiao and C.-P. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 2526–2530.
[52] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 223–227.
[53] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 1537–1540.
[54] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1089–1093.
[55] Y. Attabi and P. Dumouchel, "Anchor models for emotion recognition from speech," *IEEE Trans. Affect. Comput.*, vol. 4, no. 3, pp. 280–290, Jul. 2013.
[56] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden Markov models with deep belief networks," in *Proc. 2013 IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Olomouc, Czech Republic, Dec. 2013, pp. 216–221.
[57] P.-Y. Shih, C.-P. Chen, and H.-M. Wang, "Speech emotion recognition with skew-robust neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 2751–2755.
[58] L. Töth, "Multi-resolution spectral input for convolutional neural network-based speech recognition," in *Proc. 2017 Int. Conf. Speech Technol. Hum.-Comput. Dialogue (SpeD)*, Bucharest, Romania, Mar. 2017, pp. 1–6.

**ZIPING ZHAO** was born in Tianjin, China, in 1980. He received the B.Sc. and M.S. degrees in computer science from Tianjin Normal University, China, in 2002 and 2005, respectively, and the Ph.D. degree in automatic prediction of prosodic phrases from Nankai University, China, in 2008.

He was with Tianjin Normal University, in 2008. In 2010, he was a Visiting Scholar with the Key Laboratory of Trustworthy Computing, East China Normal University. In 2016, he became the Vice Dean of the college of computer and information engineering, Tianjin Normal University. His research interests include affective computing and machine learning.

**ZHONGTIAN BAO** was born in Ningbo, Zhejiang, China, in 1999. He received the bachelor's degree from Nanjing University, in 2017. He is currently pursuing the master's degree with Tianjin Normal University. His research interest includes speech emotion recognition and applications.

**YIQIN ZHAO** was born in Taiyuan, Shanxi, China, in 1996. He is currently pursuing the bachelor's degree in software engineering with Tianjin Normal University, where he has been a member with the Cognitive and Affective Computing Lab, since 2016. His current research interests include the intersection of affective computing, audio signal processing, and machine learning.

**ZIXING ZHANG** received the master's degree from the Beijing University of Posts and Telecommunications (BUPT), China, in 2010, and the Ph.D. degree in computer engineering from the Technical University of Munich (TUM), Germany, in 2015. He was a Postdoctoral Researcher with the University of Passau, Germany, from 2015 to 2017. He has been a Research Associate with the Department of Computing, Imperial College London (ICL), U.K., since 2017. He has authored more than 80 publications in peer-reviewed books, journals, and conference proceedings to date. His research interests include deep learning technologies for the speaker-centered state (e. g. , emotion) and health computing. He has organized special sessions, such as the IEEE 7th ACII, in 2017, and the 43nd ICASSP, in 2018, and a special issue in the IEEE TETCI, in 2019. He also serves as a Reviewer for numerous leading-in-their fields' journals and conferences and as a Programme Committee Member and the Area Chair for many international conferences.

**NICHOLAS CUMMINS** received the bachelor's degree (Hons.) and the Ph.D. degree in electrical engineering from UNSW, Australia, in 2011and 2016, respectively. He is currently pursuing the habilitation degree with the ZB.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, where he works on the Horizon 2020 projects DE-ENIGMA, RADAR-CNS, sustAGE, and TAPAS. He is an External Researcher on the National Science Foundation of China grant No.31860285, diagnosis of depression by speech signals. His PhD investigated whether the voice can be used as an objective marker in the diagnosis and monitoring of clinical depression. He has published regularly in the field of depression detection, since 2011; these papers have attracted significant attention and citations. His current research interest includes the areas of behavioral signal processing with a focus on the automatic multisensory analysis and understanding of different health states.

**ZHAO REN** received the master's degree in computer science and technology from the Northwestern Polytechnical University (NWPU), China, in 2017. She is currently pursuing the Ph.D. degree with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, where she is a Research Assistant, involved with the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Training Network project TAPAS for emotion analysis based on pathological speech. Her research interests include transfer learning, unsupervised learning, and deep learning for the application in health care and wellbeing.

**BJÖRN SCHULLER** (M'06–SM'15–F'18) received the diploma degree, in 1999, the doctoral degree in automatic speech and emotion recognition, in 2006, and the Habilitation and Adjunct Teaching Professorship in signal processing and machine intelligence, in 2012, all in electrical engineering and information technology from the Technische Universität München (TUM), Germany. He is currently a tenured Full Professor with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, and a Professor of Artificial Intelligence with the GLAM–Group on Language, Audio and Music, Department of Computing, Imperial College London, London, U.K. He is also the Honorary Dean of the International Research Centre for Affective Intelligence, Tianjin Normal University. He coauthored five books and more than 800 publications in peer-reviewed books, journals, and conference proceedings leading to more than 24 000 citations (h-index 73). He was Elected Member of the IEEE Speech and Language Processing Technical Committee, the Editor in Chief of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, the President-emeritus of the AAAC, and a Senior Member of the ACM.